How effective are current fairness intervention methods in addressing unfairness in recommendation systems, and what trade-offs do they introduce in terms of accuracy?

Jiaqing Huang | Supervisor: Masoud Mansoury

1. Background

1.1 Recommendation Systems

A **recommendation system** (RS) is an information filtering tool that predicts user preferences based on data such as user profiles, item features, and interaction history, and provides personalized item suggestions accordingly.

1.2 Fairness in recommendation systems

Fairness is a fundamental social construct and core human value that originated in philosophy, sociology, law, and economics. In the context of RS, which operate as two-sided platforms serving both users and items, fairness implies that RS should treat all users and items equitably:

- **User fairness**: whether the recommendation is fair to all users.
- **Item fairness**: whether the recommendation treats all items fairly.

Based on whether the target is to ensure individual-level or group-level fairness, fairness can be further categorized into:

- **Individual fairness**: similar individuals or items should be treated similarly.
- **Group fairness**: the protected groups should be treated similarly as the advantaged groups

1.3 Fairness intervention methods

Existing methods for improving fairness in RS can be categorized into **three types** based on the intervention stage in the recommendation pipeline (see Figure 1):

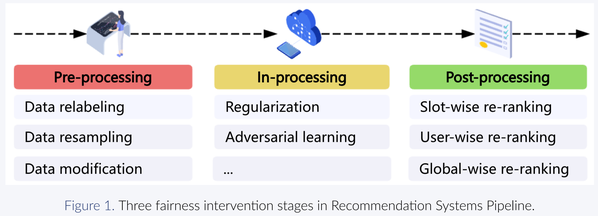


Figure 1. Three fairness intervention stages in Recommendation Systems Pipeline.

2. Research Question

How effective are current fairness intervention methods in addressing unfairness in recommendation systems, and what trade-offs do they introduce in terms of accuracy?

Sub-Questions:

- How do the current fairness intervention methods affect accuracy and fairness in RS, respectively?
- What trade-offs exist between accuracy and fairness when applying these methods to real-world datasets?
- Which type of intervention achieves the best overall balance between fairness and accuracy?

References

[1] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 12 2011. [Online]. Available: https://doi.org/10.1007/s10115-011-0463-8

[2] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 1 2018, pp. 172–186. [Online]. Available: http://proceedings.mlr.press/v81/ekstrand18b/ekstrand18b.pdf

[3] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa* ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 1569–1578.

[4] H. Steck, "Calibrated recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 154–162. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3240323.3240372

[5] A. J. Biega, K. P. Gummadi, and G. Weikum, "Equity of attention: amortizing individual fairness in rankings," 6 2018, pp. 405–414.

3. Methodology

3.1 Fairness intervention methods

Relabeling (Kamiran & Calders [1]): Modify interaction labels to equalize positive prediction rates across groups

$$P(\hat{h}_{ij} = 1 \mid a_{u_i} = a_1) \approx P(\hat{h}_{ij} = 1 \mid a_{u_i} = a_2)$$

Resampling (Ekstrand et al., [2]): Oversample or undersample interactions to balance group distributions

FA*IR (Zehlike et al., [3]): Re-ranks top-k recommendations to enforce minimum representation of protected items

$$|l^j_{u_i} \cap V_p| \geq \lfloor \alpha \cdot j \rfloor, \quad \forall j \in \{1, 2, \ldots, k\}$$

Calibration (Steck [4]): Match the distribution of recommended item categories with the user's true preference distribution

$$\mathrm{Cal}(u_i) = \mathrm{KL}(p \parallel q) = \sum_{c \in C} p(c \mid u_i) \log \frac{p(c \mid u_i)}{q(c \mid u_i)}$$

Equity of Attention (Biega et al., [5]): Promote equitable exposure by penalizing over-exposed items in ranking

$$\mathrm{score}(v_j \mid u_i) = \hat{h}_{ij} - \lambda \cdot \log(e_{v_j} + 1)$$

3.2 Evaluation Metrics

Table 1. Evaluation metrics used in this study. Arrows (↑, ↓) indicate preference direction.

Metric[1]	Interpretation
Accuracy	
`Pre@K` ↑	Proportion of recommended items that are relevant
`Rec@K` ↑	Proportion of relevant items that are retrieved
`Hit@K` ↑	At least one relevant item appears in top-K
`MAP@K` ↑	Mean precision over relevant items
`NDCG@K` ↑	Rank-sensitive relevance evaluation
User Fairness	
`UGF-NDCG@K` ↓	Accuracy gap (NDCG) between female and male user groups
`UGF-IC@K` ↓	Diversity gap (IC) between female and male user groups
Item Fairness	
`IC@K` ↑	Fraction of unique items recommended across all users
`AP@K` ↓	Mean popularity of recommended items
`SE@K` ↑	Dispersion of item exposure across users
`GI@K` ↓	Inequality in item exposure
`TP@K` ↑	Exposure to long-tail (less popular) items

[1] Metric abbreviations: `Pre` = Precision, `Rec` = Recall, `Hit` = Hit Ratio, `MAP` = Mean Average Precision, `NDCG` = Normalized Discounted Cumulative Gain, `UGF` = User Group Fairness, `IC` = Item Coverage, `AP` = Average Popularity, `SE` = Shannon Entropy, `GI` = Gini Index, `TP` = Tail Percentage.

3.3 Datasets

Dataset	Users	Items	Interactions	User-side Attributes	Item-side Attributes	Sparsity
ML-1M	6,040	3,706	1,000,209	gender, age, occupation	genres	95.532%
Lastfm-NL	8,792	36,077	434,240	gender, age	-	99.863%

Table 2. Basic information of the two datasets

3.4 Experimental Setup

Three experimental scenarios:

- Baseline Scenario
- Pre-processing Scenario
- Post-processing Scenario

The RecBole framework is used to apply the same BPR model configuration and consistent data splits across experiments.

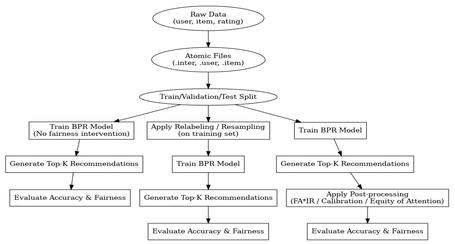


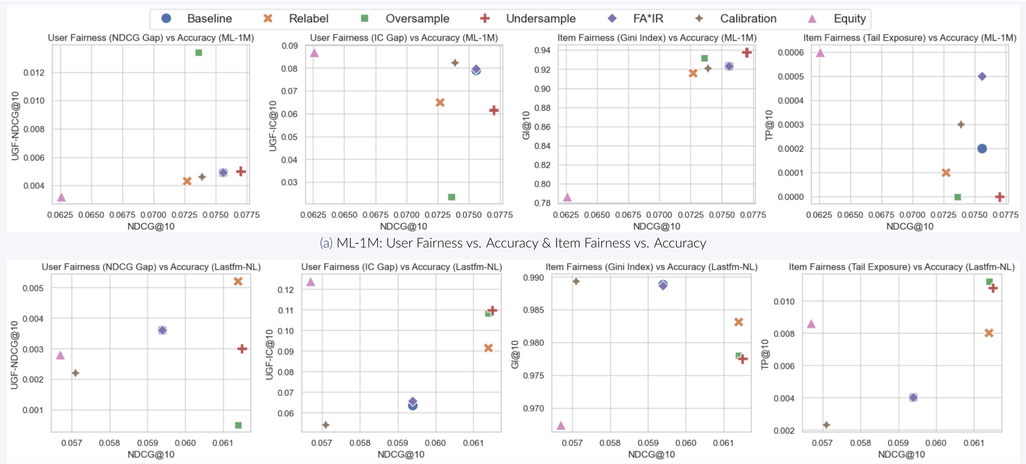
Figure 2. Experimental scenarios overview.

4. Results(1) Effects on Accuracy and Fairness

Method	Accuracy					User Fairness						Item Fairness				
	Pre↑	Rec↑	Hit↑	MAP↑	NDCG↑	NDCG(M)↑	NDCG(F)↑	UGF-NDCG↓	IC(M)↑	IC(F)↑	UGF-IC↓	IC↑	AP↓	SE↑	GI↓	TP↑
ML-1M dataset																
`Baseline`	0.0547	0.0787	0.4283	0.0326	0.0756	0.0770	0.0721	0.0049	0.3527	0.2737	0.0790	0.3758	1299.8182	0.7930	0.9230	0.0002
`Relabel`	0.0535	0.0772	0.4210	0.0306	0.0727	0.0739	0.0696	0.0043	0.3641	0.2992	0.0649	0.3986	1276.9728	0.7976	0.9158	0.0001
`Oversample`	0.0542	0.0759	0.4238	0.0313	0.0736	0.0774	0.0640	0.0134	0.3082	0.2846	**0.0236**	0.3554	1358.2146	0.7800	0.9319	0.0000
`Undersample`	**0.0565**	**0.0800**	**0.4361**	**0.0331**	**0.0770**	**0.0784**	**0.0734**	0.0050	0.3055	0.2438	0.0616	0.3323	1392.1854	0.7740	0.9377	0.0000
`FA*IR`	0.0547	0.0787	0.4283	0.0326	0.0756	0.0770	0.0721	0.0049	0.3543	0.2748	0.0796	0.3785	1299.7498	0.7923	0.9230	0.0005
`Calibration`	0.0529	0.0762	0.4200	0.0320	0.0739	0.0752	0.0706	0.0046	0.3562	0.2740	0.0823	0.3839	1237.6068	0.7992	0.9208	0.0003
`Equity`	0.0454	0.0660	0.3719	0.0266	0.0626	0.0635	0.0603	**0.0032**	**0.5468**	**0.4600**	0.0869	**0.5778**	**842.5324**	**0.8975**	**0.7863**	**0.0006**
Lastfm-NL dataset																
`Baseline`	0.0332	0.0753	0.2902	0.0270	0.0594	0.0601	0.0565	0.0036	0.1016	0.0381	0.0635	0.1116	631.4422	0.6938	0.9888	0.0040
`Relabel`	0.0345	0.0787	0.2981	**0.0279**	0.0614	**0.0624**	0.0572	0.0052	0.1404	0.0490	0.0914	0.1575	599.8631	0.6912	0.9831	0.0080
`Oversample`	**0.0350**	**0.0797**	**0.3049**	0.0275	0.0614	0.0613	**0.0618**	**0.0005**	0.1650	0.0568	0.1082	0.1860	537.2037	0.7136	0.9780	**0.0112**
`Undersample`	0.0347	0.0786	0.2974	**0.0279**	**0.0615**	0.0621	0.0591	0.0030	0.1674	0.0577	0.1097	0.1895	553.2270	0.7099	0.9775	0.0108
`FA*IR`	0.0332	0.0753	0.2902	0.0270	0.0594	0.0601	0.0565	0.0036	0.1042	0.0386	0.0656	0.1147	631.1572	0.6922	0.9886	0.0040
`Calibration`	0.0312	0.0711	0.2762	0.0263	0.0571	0.0576	0.0554	0.0022	0.0920	0.0379	**0.0541**	0.1011	590.1089	0.7074	0.9893	0.0023
`Equity`	0.0323	0.0735	0.2804	0.0254	0.0567	0.0573	0.0545	0.0028	**0.1951**	**0.0714**	0.1237	**0.2159**	**487.5218**	**0.7500**	**0.9674**	0.0086

Table 3. Performance comparison of fairness intervention methods on the ML-1M and Lastfm-NL datasets. Accuracy metrics include Pre@10, Rec@10, Hit@10, MAP@10, and NDCG@10. Fairness metrics cover user-side (UGF-NDCG@10, UGF-IC@10) and item-side (IC@10, AP@10, SE@10, GI@10, TP@10).

4. Results(2) Trade-offs Between Accuracy and Fairness



(a) ML-1M: User Fairness vs. Accuracy & Item Fairness vs. Accuracy

(b) Lastfm-NL: User Fairness vs. Accuracy & Item Fairness vs. Accuracy

Figure 3. Trade-offs between accuracy and fairness across different intervention methods on the ML-1M and Lastfm-NL datasets.

5. Conclusions and Future Work

Conclusions

No single method is optimal for all goals:

- **Accuracy-focused:** *Undersample* improves accuracy but has fairness variance; *FA*IR* slightly improves fairness with no accuracy loss.
- **User fairness-focused:** *Oversample* achieves strong gains in group fairness and diversity.
- **Item fairness-focused:** *Equity* improves long-tail exposure and balance, with some accuracy loss.

Future Work

- **In-processing methods:** Integrate fairness into model training.
- **Hybrid approaches:** Combine pre- and post-processing interventions.
- **Complex attributes:** Apply to other domains such as employment, healthcare, education.
- **Intersectionality:** Explore fairness across combined attributes (e.g., gender × age).
- **Parameter tuning:** Adapt methods to real-world system needs.